ADDITIVE BAYESIAN MODEL MEETS OBSERVATIONAL DATA

University of Zurich UZH

GILLES KRATZER, APPLIED STATISTICS GROUP, UZH

SONJA HARTNACK, VETSUISSE, UZH

http://r-bayesian-networks.org/

gilles.kratzer@math.uzh.ch

sonja.hartnack@access.uzh.ch

ECVPH WORKSHOP, ZURICH 7-9 MAY 2019

Credit Card Fraud Detection Using Bayesian and Neural Networks

Sam Maes Karl Tuyls Bram Vanschoenwinkel
Bernard Manderick
Vrije Universiteit Brussel - Department of Computer Science
Computational Modeling Lab (COMO)
Pleinlaan 2
B-1050 Brussel, Belgium
{sammaes@,ktuyls@,bvschoen@,bernard@arti.}vub.ac.be

Abstract

This paper discusses automated credit card fraud detection by means of machine learning. In an era of digitalization, credit card fraud detection is of great importance to financial institutions. We apply two machine learning techniques suited for reasoning under uncertainty: artificial neural networks and

do the fraud detection. After a process of learning, the program is supposed to be able to correctly classify a transaction it has never seen before as fraudulent or not fraudulent, given some features of that transaction.

The structure of this paper is as follows: first we introduce the reader to the domain of credit card fraud detection. In Sections 3 and 4 we briefly ex-

MOTIVATIONAL EXAMPLE: CREDIT CARD FRAUD DETECTION PREDICTION

Credit Card Fraud Detection Using Bayesian and Neural Networks

Sam Maes Karl Tuyls Bram Vanschoenwinkel

experiment	±10% false pos	±15% false pos
ANN-fig 2(a)	60% true pos	70% true pos
ANN-fig 2(a)	47% true pos	58% true pos
ANN-fig 2(c)	60% true pos	70% true pos
BBN-fig 2(e)	68% true pos	74% true pos
BBN-fig 2(g)	68% true pos	74% true pos

Table 1: This table compares the results achieved with ANN and BBN, for a false positive rate of respectively 10% and 15%.

Abstract

This paper discusses
tection by means of
of digitalization, cre
great importance to
two machine learning techniques suited for reasoning under uncertainty: artificial neural networks and

rocess of learning,
e to correctly clas-
n before as fraud-
e features of that

s follows: first we introduce the reader to the domain of credit card fraud detection. In Sections 3 and 4 we briefly ex-

MOTIVATIONAL EXAMPLE: VETERINARY EPIDEMIOLOGY DATA VISUALISATION

ELSEVIER

Contents lists available at SciVerse ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Using Bayesian networks to explore the role of weather as a potential determinant of disease in pigs

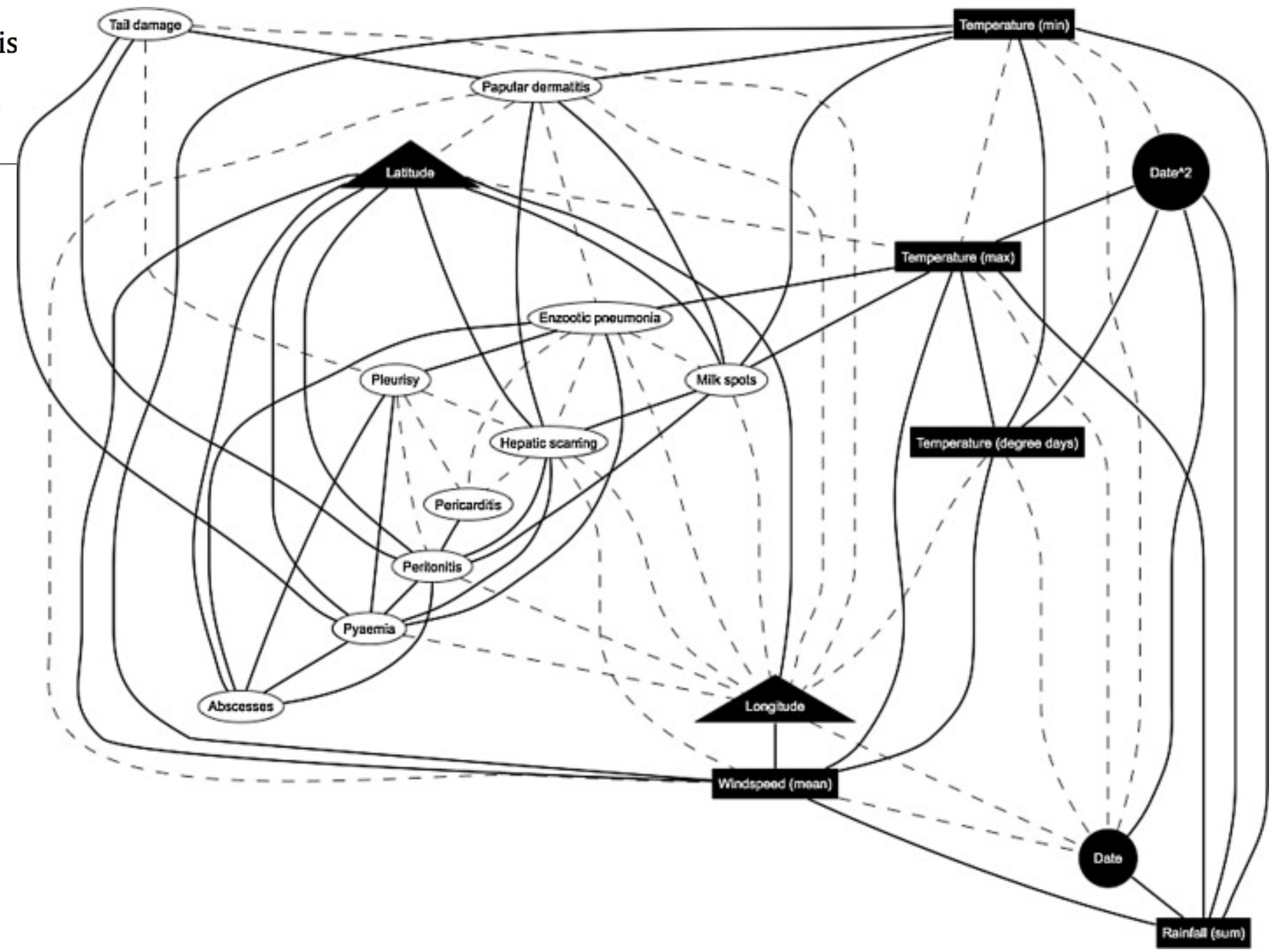
CrossMark

B.J.J. McCormick[a], M.J. Sanchez-Vazquez[b], F.I. Lewis

[a] Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

[b] OIE Organisation Mondiale de la Santé Animale, 12, rue de Prony, 75017 Paris, France

[c] Section of Epidemiology, University of Zurich, Zurich, Switzerland



Discovering complex interrelationships between socioeconomic status and health in Europe: A case study applying Bayesian Networks

Javier Alvarez-Galvez [a, b, *]

[a] Loyola University Andalusia, Department of International Studies, Campus de Palmas Altas, Faculty of Political Sciences and Law, Seville 41014, Spain
[b] Complutense University of Madrid, Department of Sociology IV (Research Methodology and Communication Theory), Campus de Somosaguas, Faculty of Political

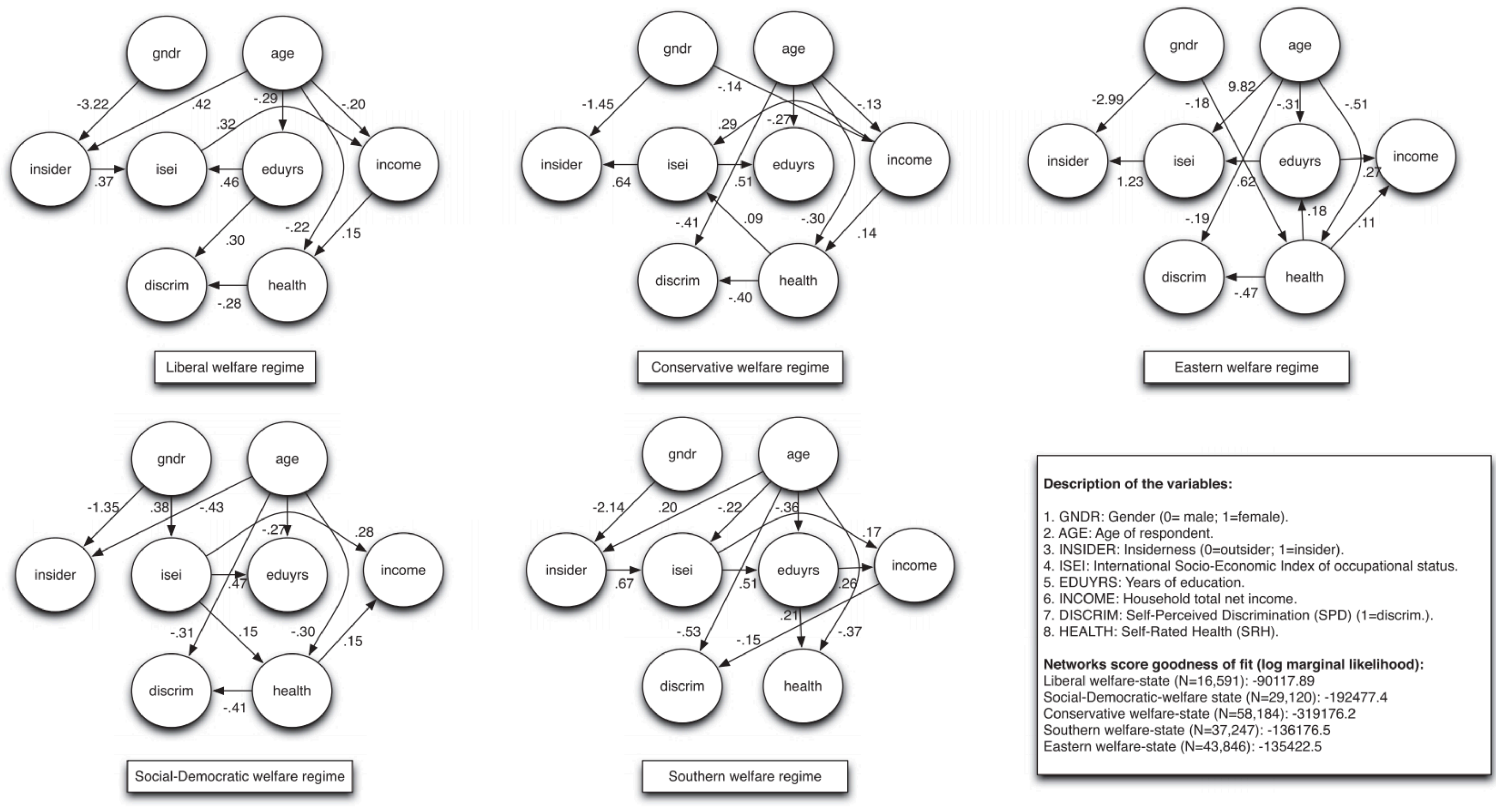


Fig. 1. Bayesian networks describing interrelationships between SES and health in five European welfare states.

BAYESIAN NETWORKS IN THE MACHINE LEARNING WORLD

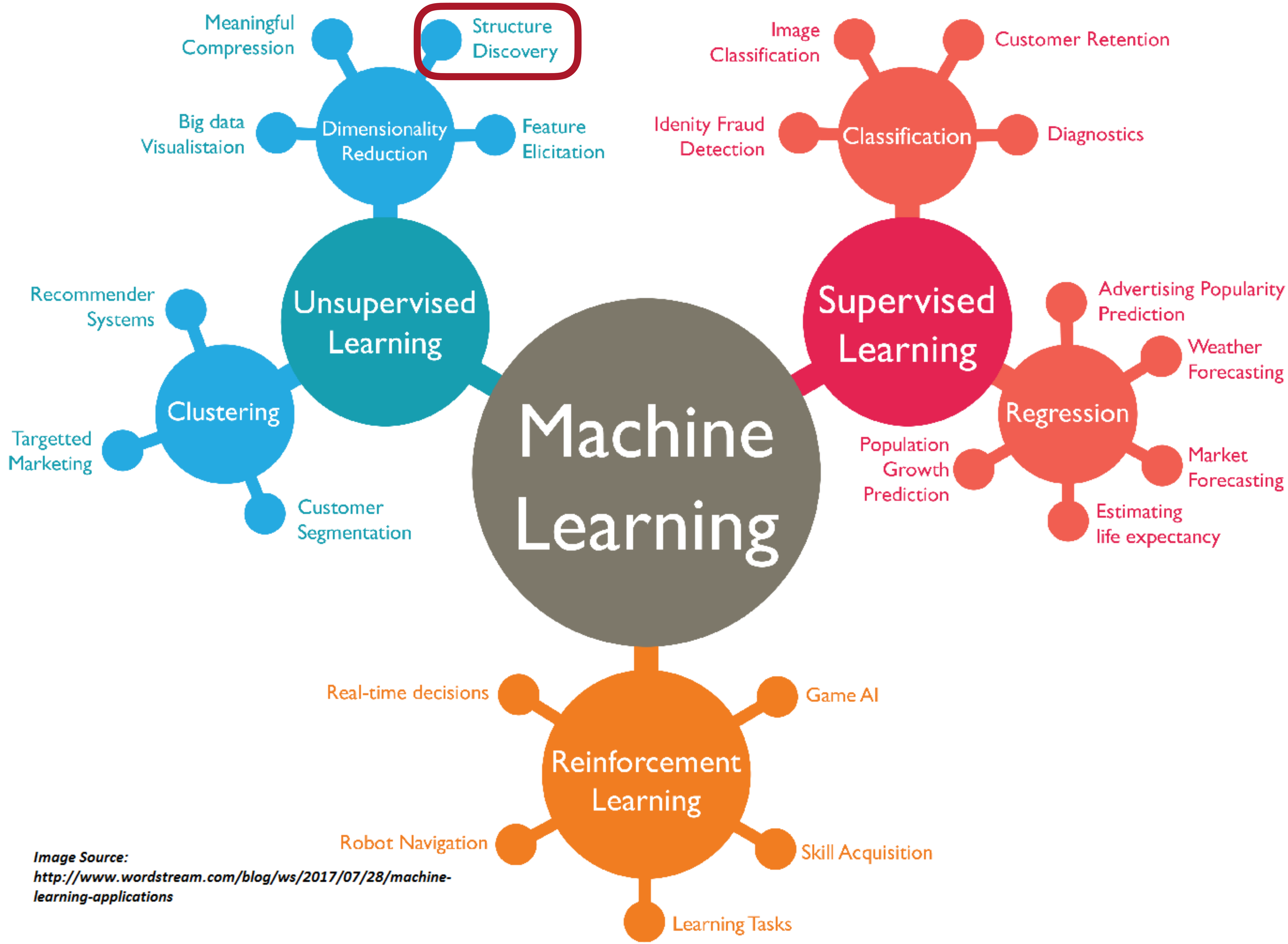


Image Source: http://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Objective of the workshop:

How to learn Bayesian networks from observational data?

Objectif of the workshop:

select

How to ~~learn~~ Bayesian networks from observational data?

Bayesian Networks are defined by two elements:

Network structure:

Directed Acyclic Graph (DAG): G = (V, A)

in which each node vi ∊ V corresponds to a random variable Xi

Probability distribution:

Probability distribution X with parameters Θ, which can be factorised into smaller local probability distributions according to the arcs aij ∊ A present in the graph.

A BN encodes the factorisation of the joint distribution

$$P(\mathbf{X}) = \prod_{j=1}^{n} P(X_j \mid \mathbf{Pa}_j, \Theta_j), \text{ where } \mathbf{Pa}_j \text{ is the set of parents of } X_j$$

PLAN

1. From observationnal dataset deduce probabilistic model
    - Usually discrete BN or jointly Gaussian
    - Epidemiological constrain: mixture of distributions
2. From probabilistic model deduce structure

EXPONENTIAL FAMILY

Observational dataset

X1	X2	X3	…
12	23	53	…
32	31	23	…
10	16	45	…
…	…	…	…

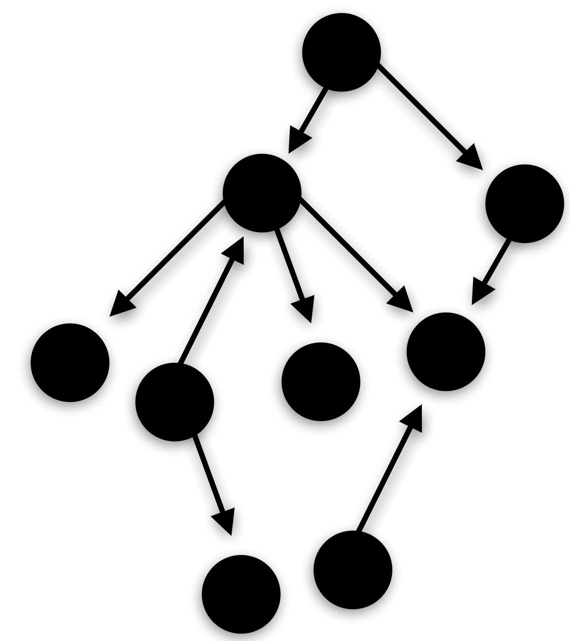
1

Probabilistic model

$$P(X_1, \ldots, X_n) = P(X_i | X_j, \ldots) \ldots$$

2

Network structure



COMBINATORIAL WALL

# Nodes	# DAGs	Inference
1 - 15 Nodes	< 10^{41} DAGs	Exact inference
16 - 25 Nodes	< 10^{100} DAGs	Exact inference possible
26 - 50 Nodes	< 10^{400} DAGs	Approximate inference
51 - 100 Nodes	< 10^{1700} DAGs	Approximate inference
101 - 1000 Nodes	< 10^{100000} DAGs	(very) approximative inference

Typical domain of interest

Approximations:

- limiting number of parents per node
- Decomposable scores/efficient algorithm
- Score equivalence

PLAN

1. From observationnal dataset deduce probabilistic model
    - Usually discrete BN or jointly Gaussian
    - Epidemiological constrain: mixture of distributions
2. From probabilistic model deduce structure

EXPONENTIAL FAMILY

Observational dataset

X1	X2	X3	...
12	23	53	...
32	31	23	...
10	16	45	...
...	...	...	...

1

Probabilistic model

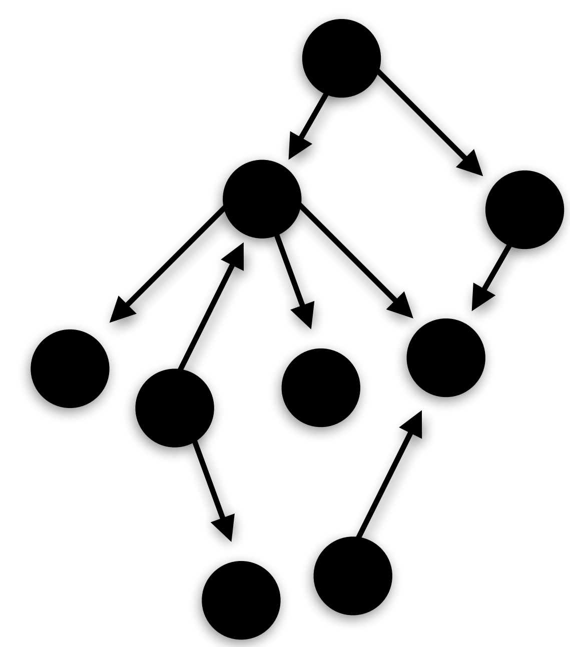
$$P(X_1, \dots, X_n) = P(X_i | X_j, \dots) \dots$$

Independance testing

Computing directly

2

Network structure



SOME ELEMENTS OF PROBABILITY THEORY

The conditional probability of A given B is: $$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

Bayes theorem: $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Let A, B and C non intersecting subsets of nodes in a DAG G

A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B | C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

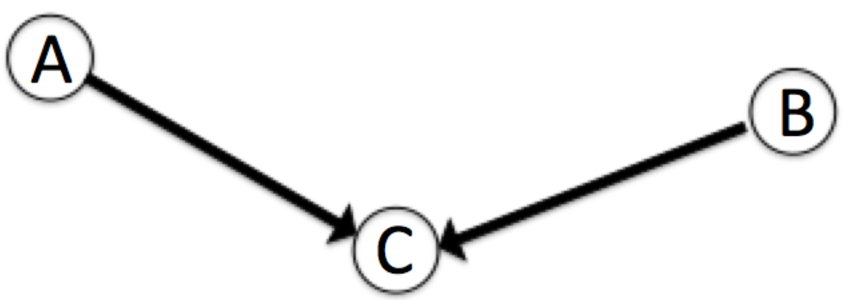
ELEMENT OF GRAPH THEORY

Let A, B and C non intersecting subsets of nodes in a DAG G

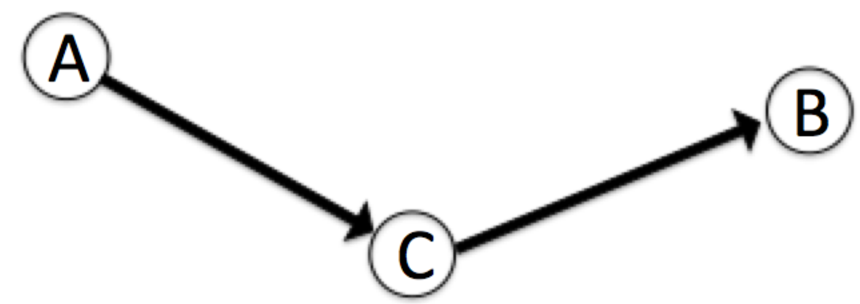
A is conditionally independent of B given C if: $A \perp\!\!\!\perp_P B|C$

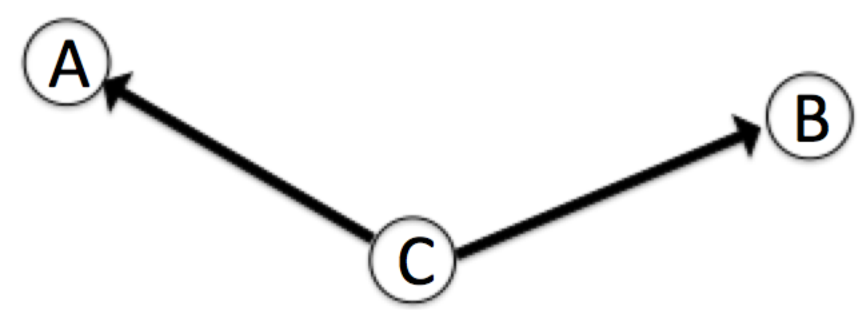
$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

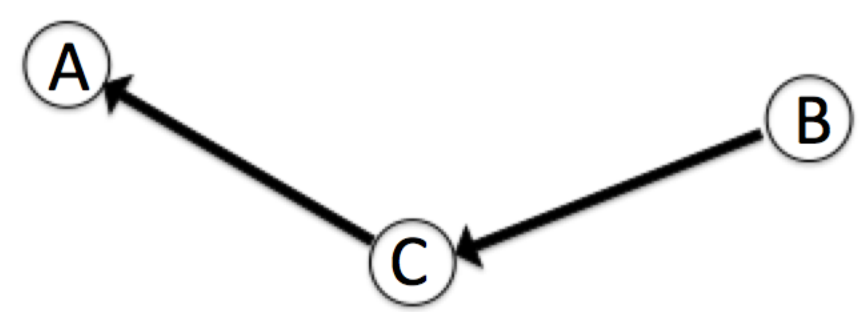
$$A \not\perp\!\!\!\perp_P B|C$$



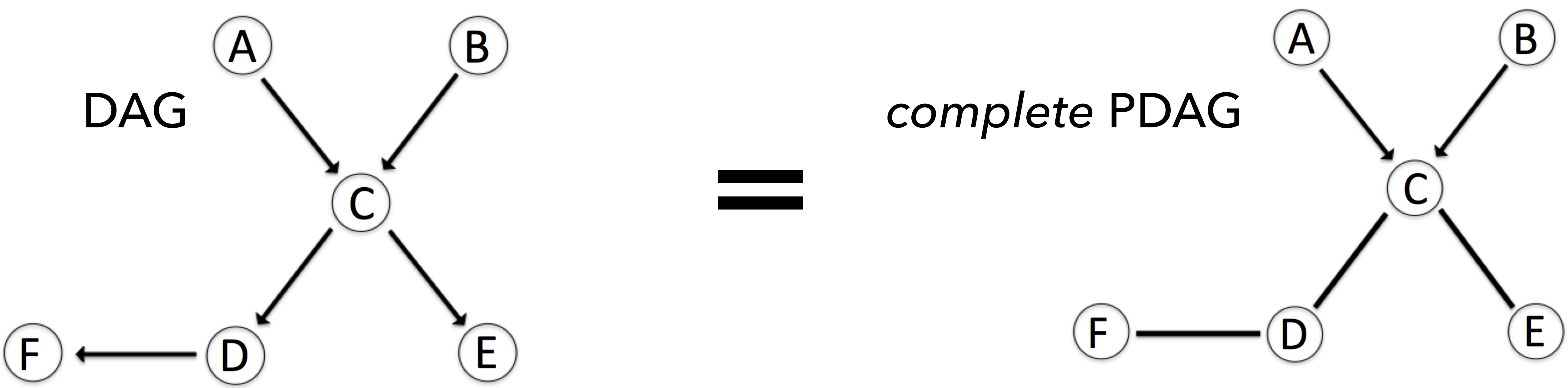
$$A \perp\!\!\!\perp_P B|C$$





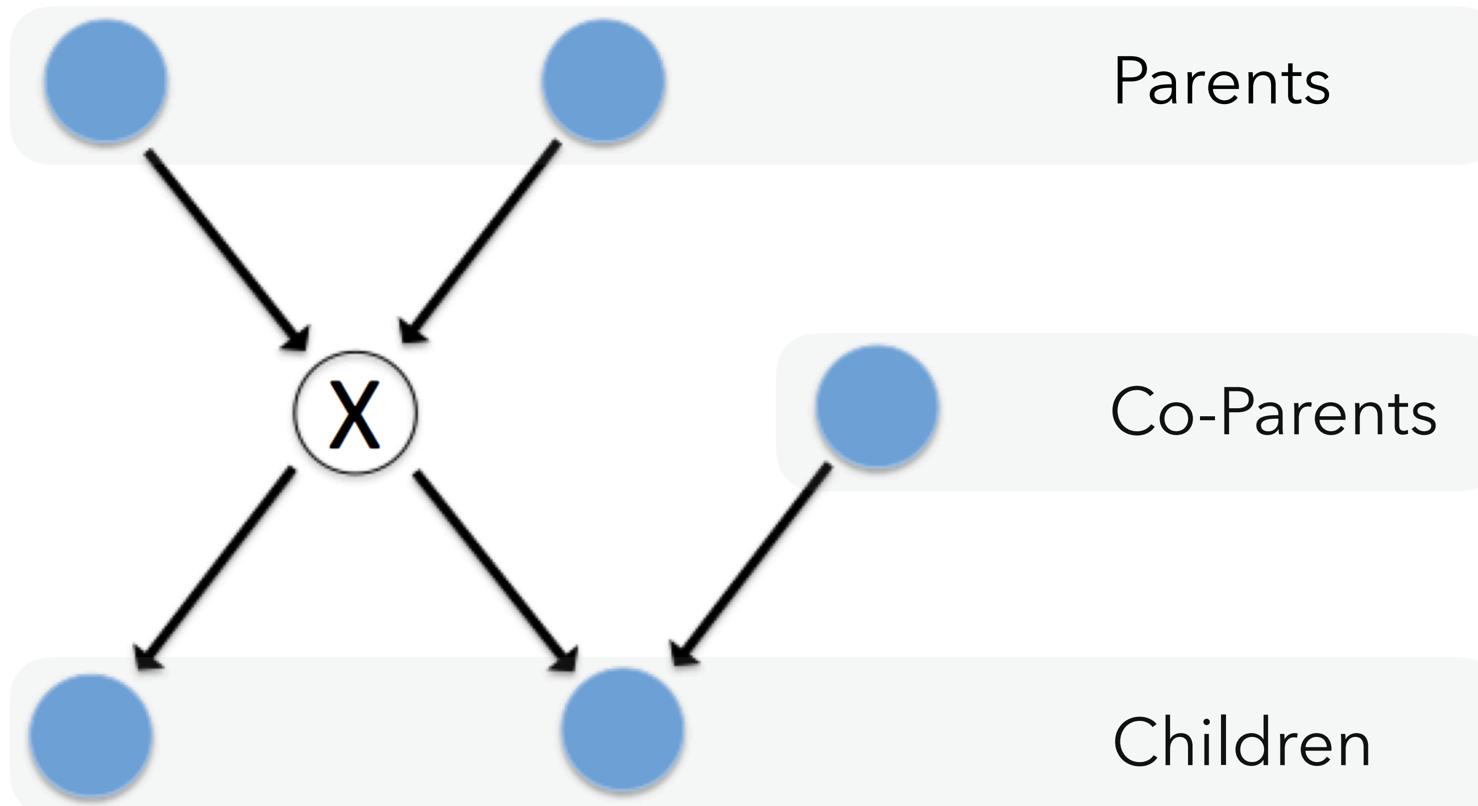


- In a practical perspective, for observational data, if learning algorithms rely on probabilistic learning algorithm. Then one can learn up to the Markov equivalence class.
- Markov equivalence class are the set of DAGs that have the same skeleton and v-structure.



ELEMENT OF GRAPH THEORY: MARKOV BLANKET

The Markov Blanket of a node is the set of parents, co-parents and children.



$$P(X_k \mid X_n, k \neq n) = P(X_k \mid X_{\mathrm{MB}(k)}), \forall k$$

The Markov Blanket of a node is the set of nodes that shields the index node from the rest of the network

Local Markov property:

$$X \perp \text{ Non-Descendants(X) } | Pa(X)$$

$$\mathcal{M} = (\mathcal{S}, \Theta_{\mathcal{M}})$$

Model selection

Structure learning

Parameter estimation

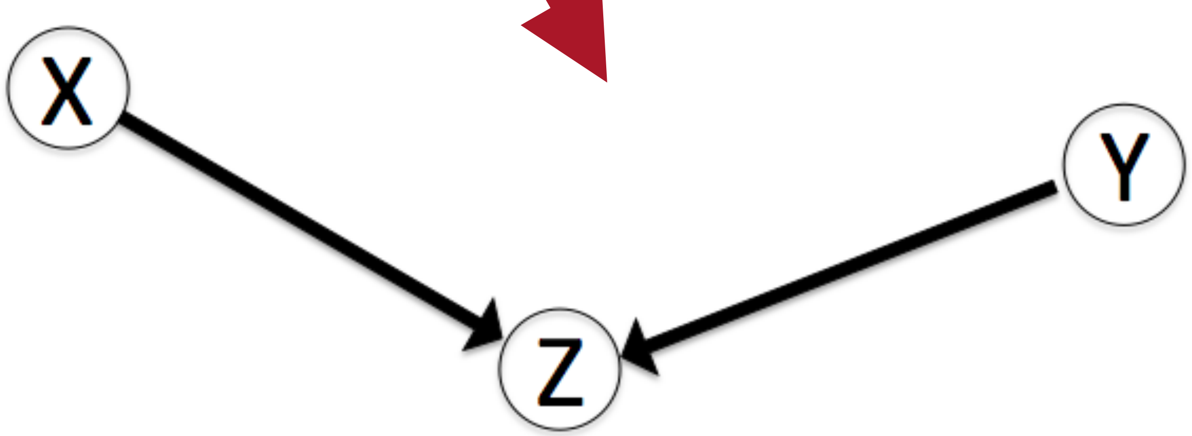
Parameter learning

$$P(\mathcal{M}|\mathcal{D}) = \underbrace{P(\Theta_{\mathcal{M}}, \mathcal{S}|\mathcal{D})}_{\text{model learning}} = \underbrace{P(\Theta_{\mathcal{M}}|\mathcal{S}, \mathcal{D})}_{\text{parameter learning}} \cdot \underbrace{P(\mathcal{S}|\mathcal{D})}_{\text{structure learning}}$$

Constraint based algorithms

$$P_{X \perp\!\!\!\perp Y | Z} < \alpha$$

$$X \perp\!\!\!\perp_{S} Y|Z = X \perp Y|Z$$



Search-and-score algorithms

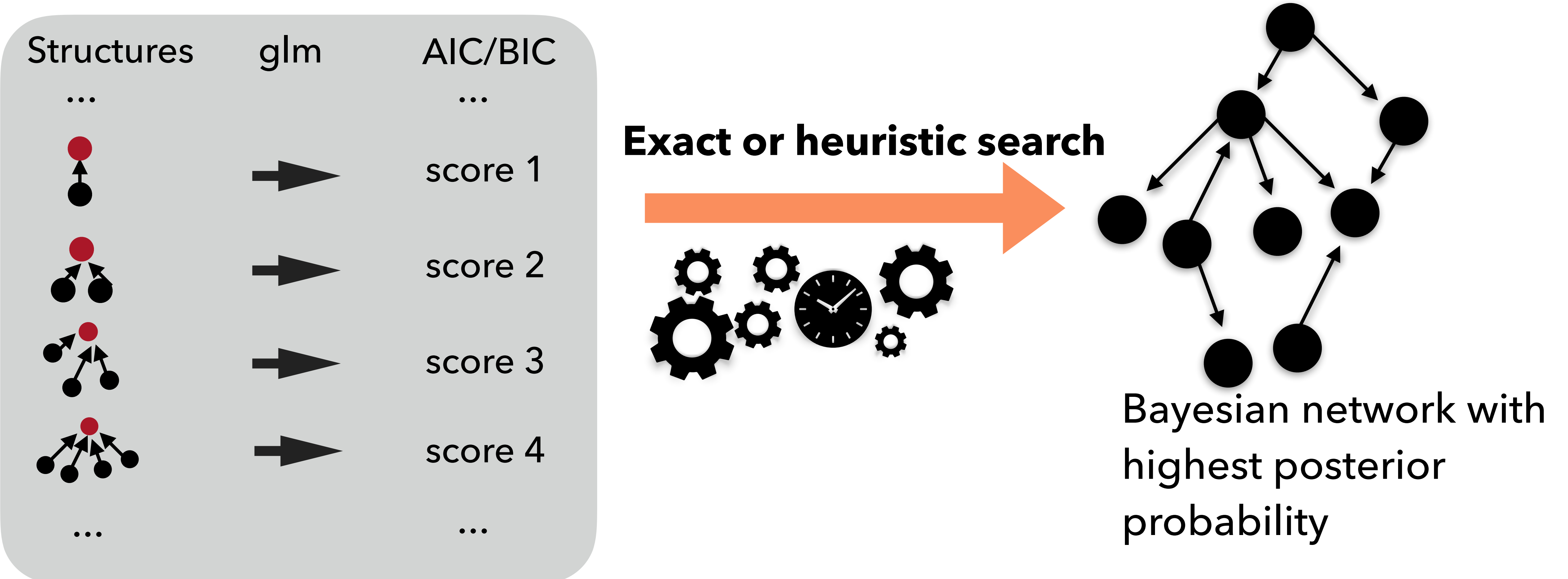
Maximum a posteriori score

$$G^* = \underset{G}{\operatorname{argmax}} \; f(\mathcal{D}, G, n, \dots)$$

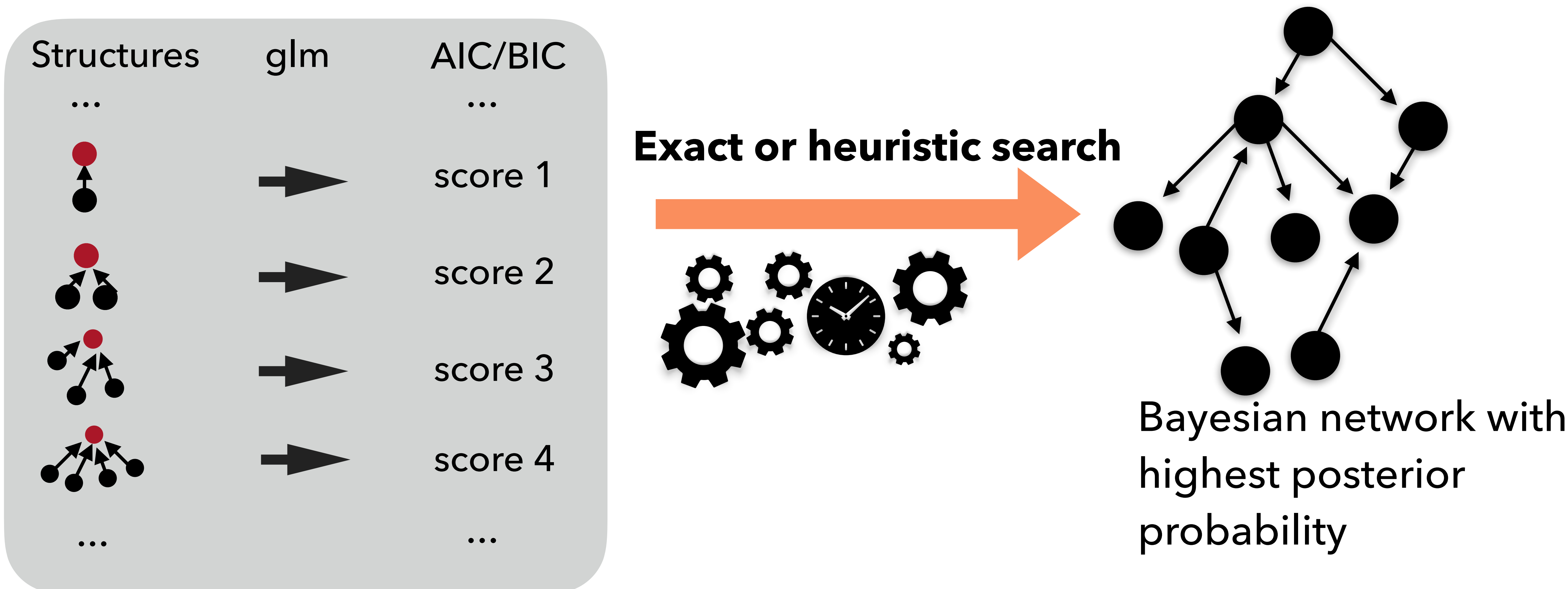
Example of scoring functions:

- Bayesian or ML scores
  - Bayesian Posterior
  - Bayesian-Dirichlet (BDeu,BDs,BDe)
  - Bayesian Information Criterion (BIC)

Search and score algorithm

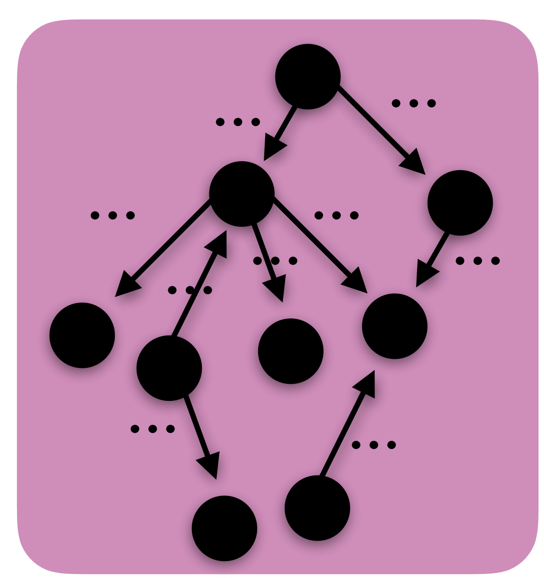


Search and score algorithm

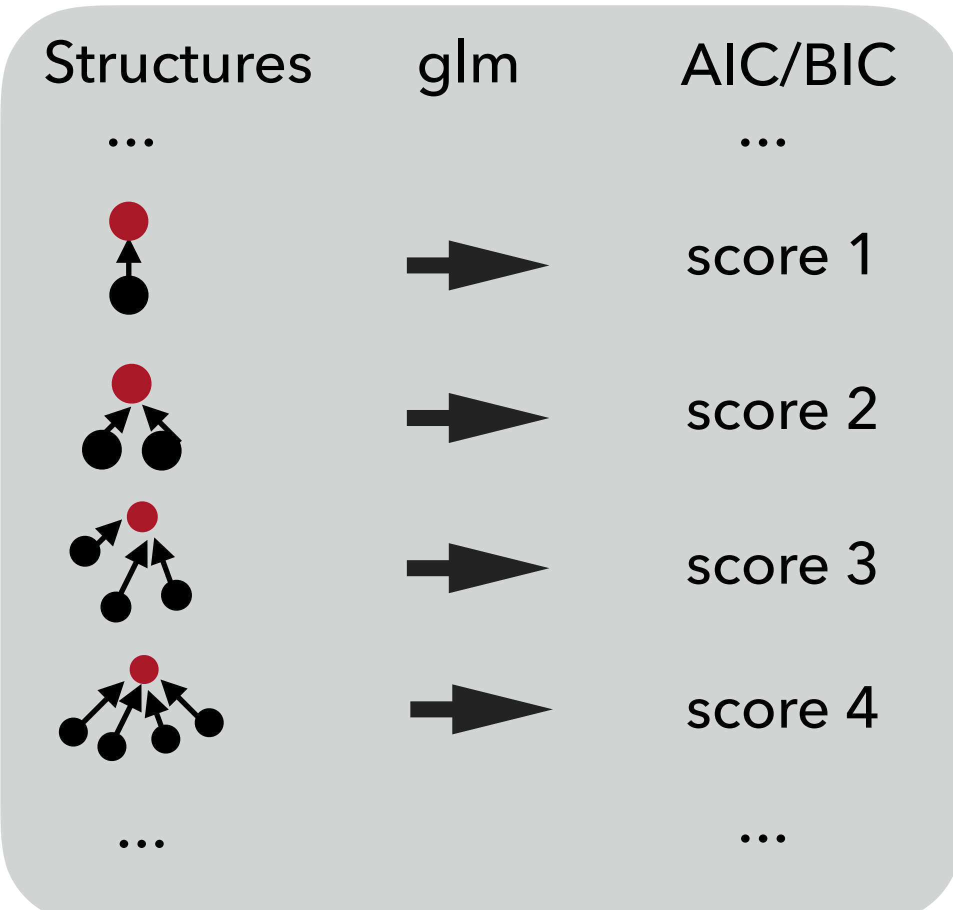


Parameter estimation

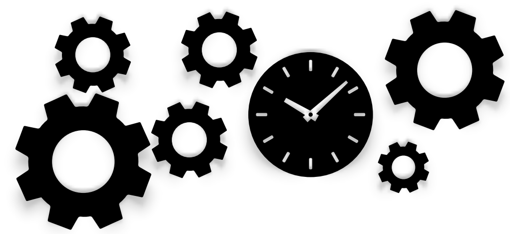
- compute marginal posterior density
- regression estimate



Search and score algorithm

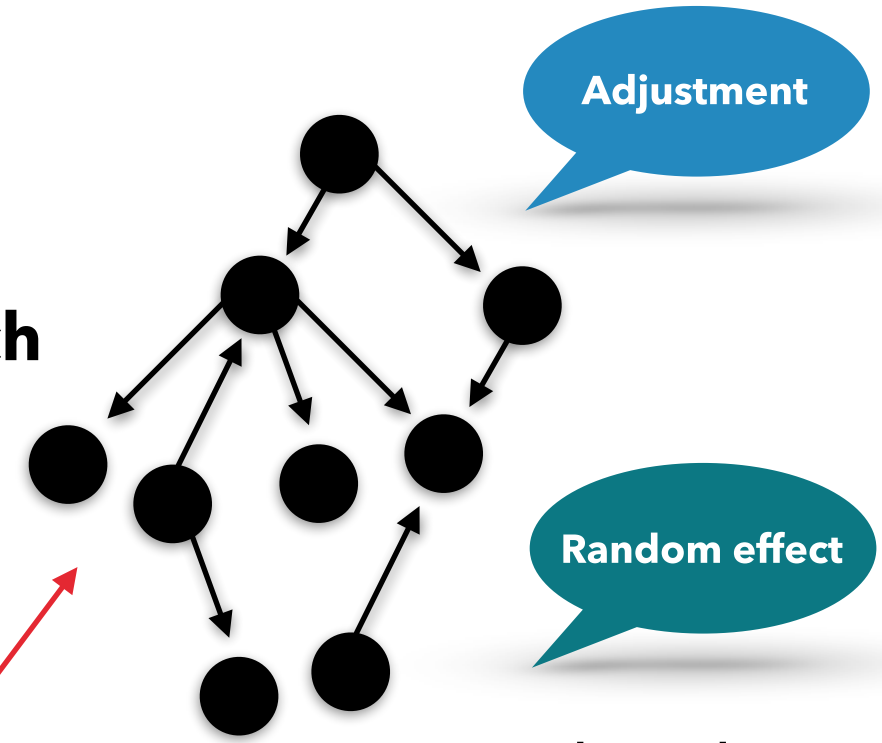


Exact or heuristic search



Causality!

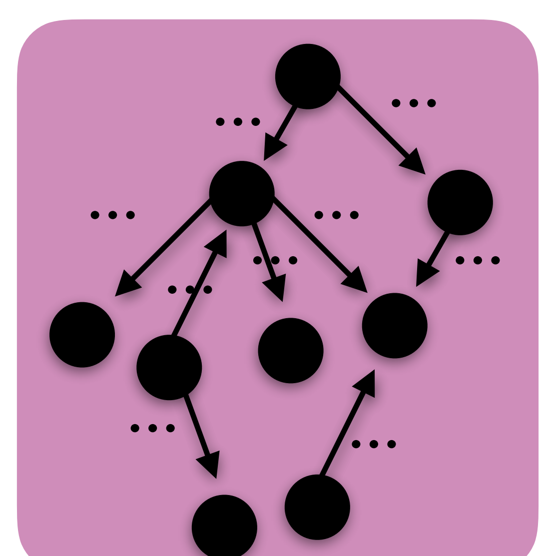
Ban/Retain structures



Bayesian network with highest posterior probability

Parameter estimation

- compute marginal posterior density
- regression estimate



Using R

buildscorecache()

mostprobable()

fitabn()

R CODE: SOFTWARE IMPLEMENTATION

Popular R packages (available on CRAN)

bnlearn

- Learning via constraint-based and score-based algorithms (many!)

pcalg

- Robust estimation of CPDAG via the PC-Algorithm

deal

- Learning BNs with mixed (discrete and continuous) variables

catnet

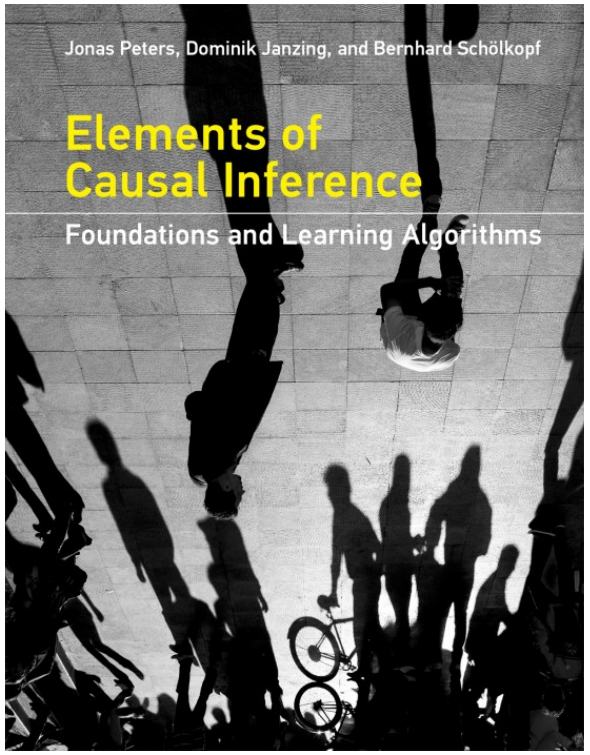
- Discrete BNs using likelihood-based criteria

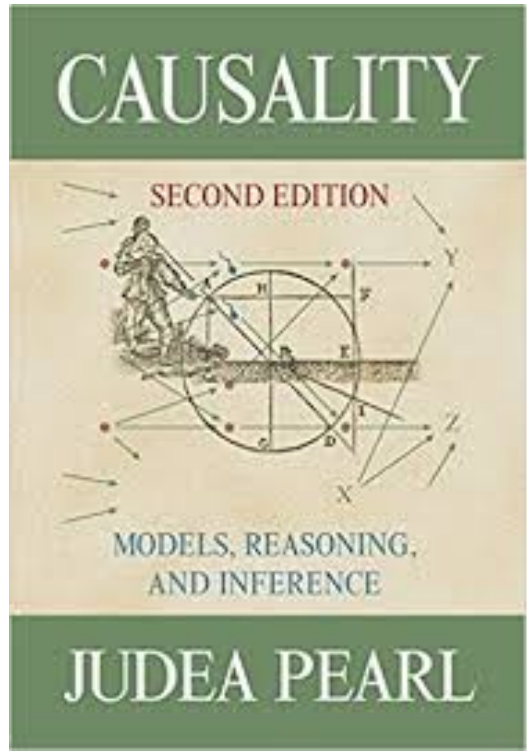
abn

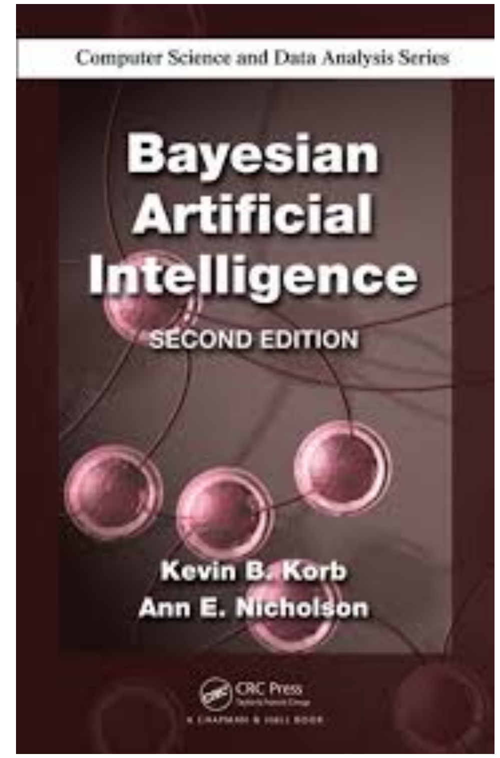
- Learning BNs with mixed (discrete, continuous, Poisson) variables
- Score based methods: Bayesian and frequentist estimation
- Exact and heuristic search
- Link strength

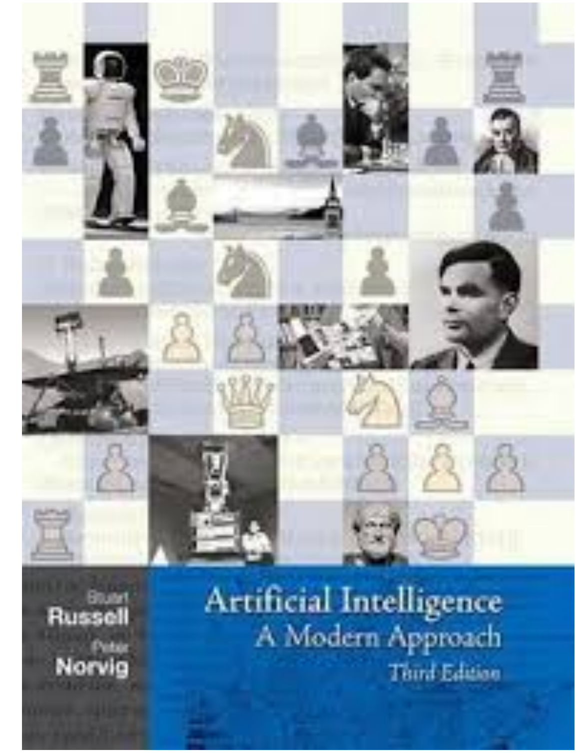
Disclaimer: I am author and maintainer of the abn R package. I will use it for the example part.

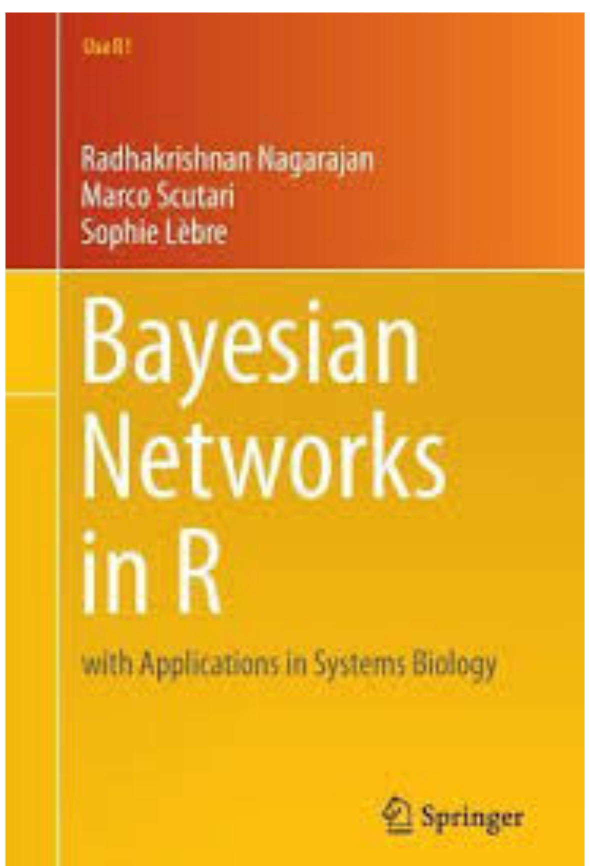
SELECTED BIBLIOGRAPHY

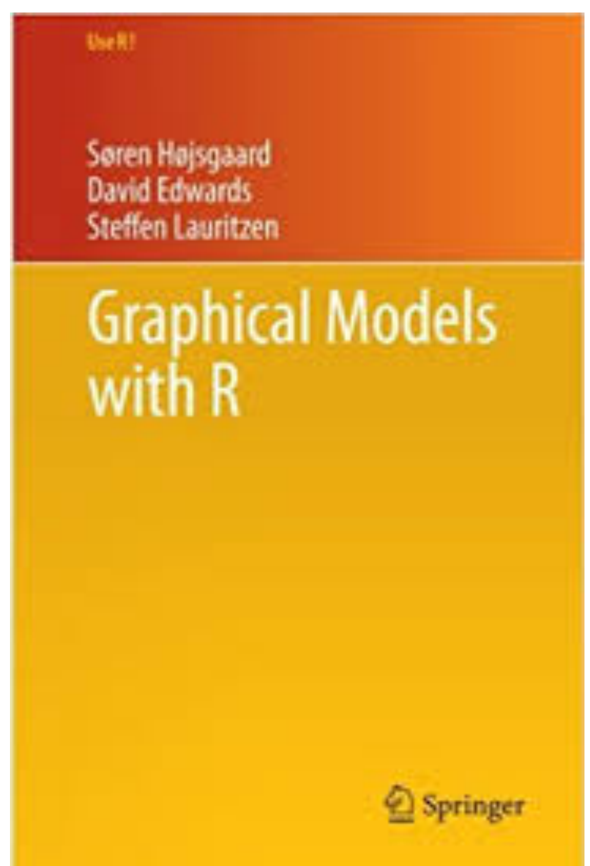


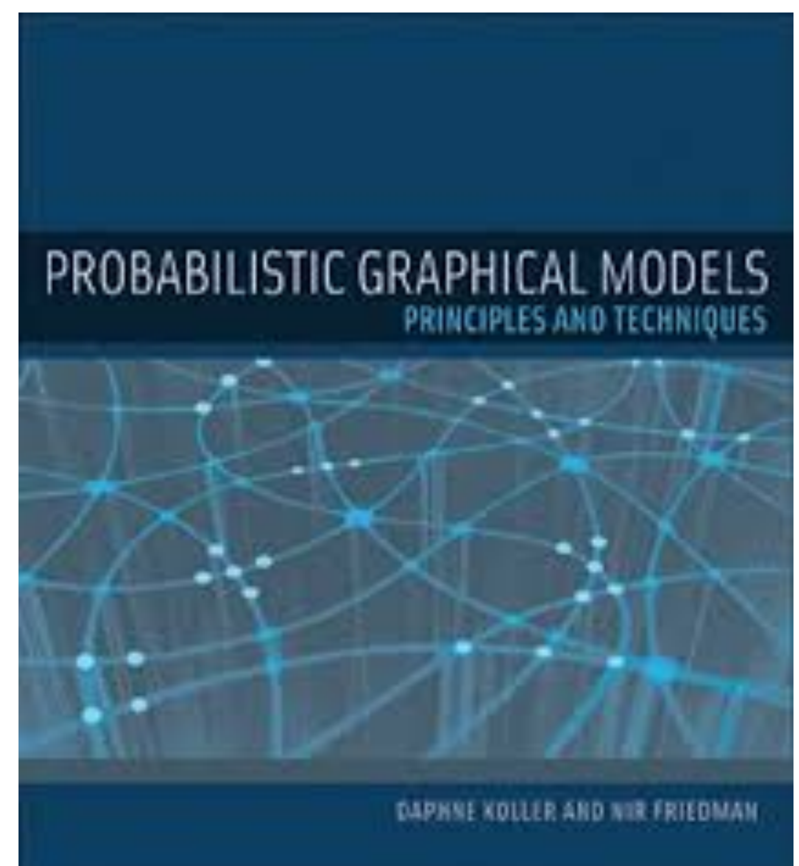




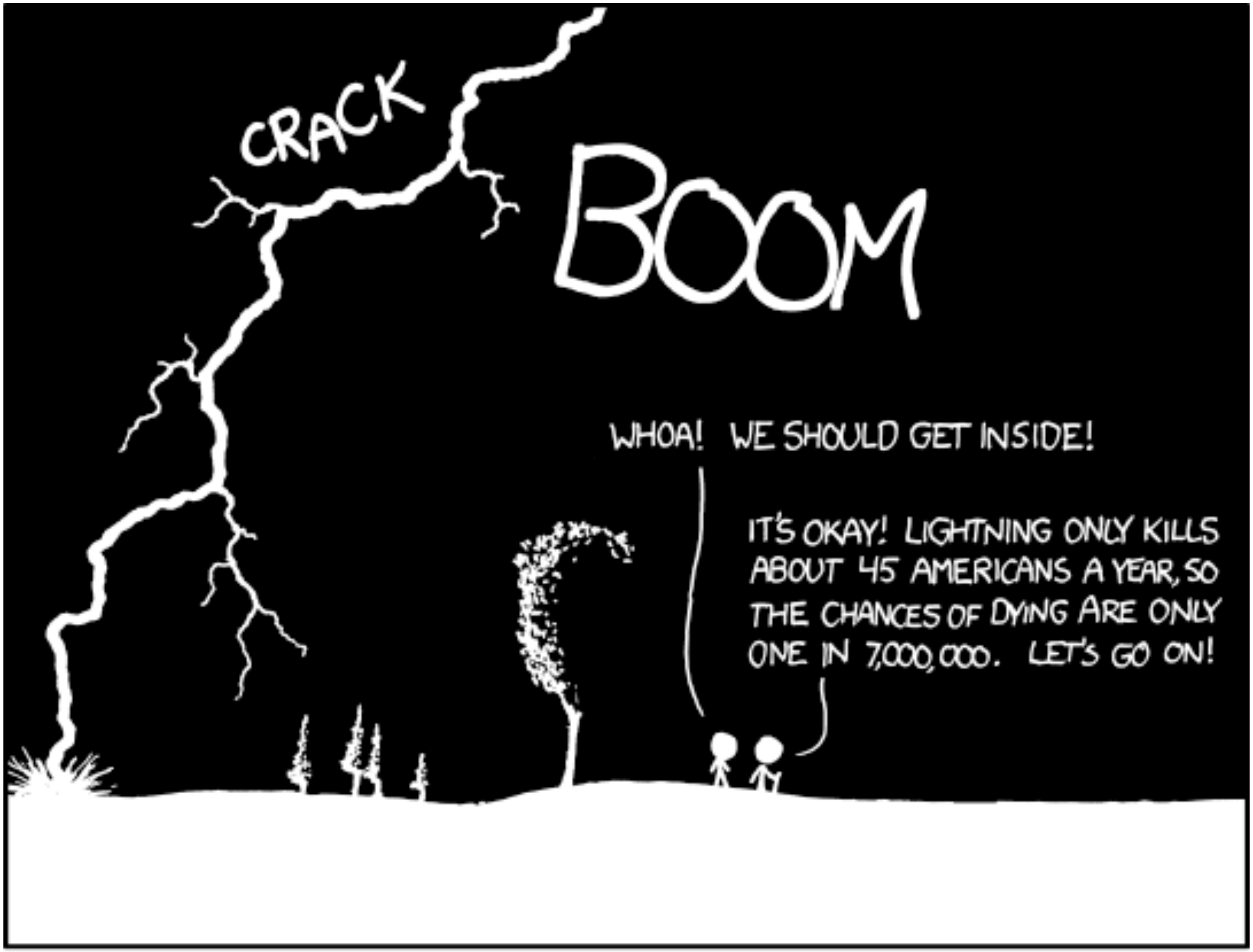








Thank you for your attention



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

xkcd.com